A promoter trap in embryonic stem (ES) cells selects for integration of DNA into CpG islands

A promoter trap in embryonic stem (ES) cells selects for integration of DNA into CpG islands

Donald Macleod*, Robin Lovell-Badge[1], Sinead Jones and Ian Jackson
MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU and [1]Laboratory of Eukaryotic Molecular Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Received October 15, 1990; Revised and Accepted November 26, 1990 EMBL accession nos X54811, X54812

ABSTRACT

An analysis of several G418-resistant ES cell lines produced by electroporation of a promoterless *neo*^{R} gene (NASTI), shows an enrichment for integrations within, or adjacent to, CpG islands. A detailed analysis of two of the cell lines reveals short regions of homology between the genomic target DNA and the construct ends, and that recombination may be mediated by DNA Topoisomerase I. The DNA flanking the insert detects transcription of endogenous genes, and in one cell line divergent transcripts are detected. This use of ES cells should provide an effective and efficient means of creating insertional mutations in mice.

INTRODUCTION

Embryonic stem (ES) cells are pluripotent stem cells derived from the mouse blastocyst inner cell mass. Cell lines can be maintained in an undifferentiated state in culture (1,2), where they can be genetically manipulated and characterized. The ES cells can then go on to form chimaeras by microinjection into host mouse blastocysts. The mosaic of host and ES cells produces a chimaeric animal in which the ES cells are able to contribute to all cell types (3). Most importantly the manipulated ES cells are able to contribute to the germ line through which the phenotype of a mutation, which had been previously characterized *in vitro*, can be assessed in succeeding generations in both heterozygous and homozygous form. The use of ES cells has received much attention recently as a means of gene targetting by homologous recombination (4).

We are interested in identifying, cloning and mutagenizing genes involved in the early stages of development and morphogenesis in the mouse, and genes which are active in ES cells will include some of these. We have therefore developed a method by which random genes active in ES cells can be identified and mutated by means of a promoter-trap. The construct used for transfection is essentially a promoterless *neo*^{R} gene. Only cells in which the insertion has occurred in the correct orientation next to an active promoter will be able to express the *neo*^{R} gene and will therefore survive G418 selection. The cloning of the DNA flanking the insertion sites allows a detailed analysis of the integration sites of these non-homologous or 'illegitimate' recombination events.

Several cell lines have been produced from G418 resistant foci, resulting from a single transfection experiment, and in each of these the transfected DNA has integrated into different regions of the mouse genome. Analysis of cell line DNA reveals that in six of seven lines examined the integrated DNA has inserted close to or within a CpG island. Such islands are found at the 5′ end of all housekeeping genes and some tissue specific genes (5,6) and are associated with a different chromatin structure (7).

The precise genomic integration sites in two of the cell lines have been analysed in detail, and there are homologies to consensus recognition sites for DNA topoisomerase I at the breakpoints. There are also short regions of homology between the target DNA and the construct ends. We discuss the possibility that DNA topoisomerase I may be involved in mediating the recombination events in these cell lines.

EXPERIMENTAL METHODS

Cell culture and electroporation

In the original transfection experiment from which the cell lines were derived, CCE ES cells (8), were electroporated with the linearized DNA construct NASTI, at 0.01 mg/ml, using a wokshop-made apparatus (9). ES cells were cultured using BRL conditioned media and mitomycin-treated, STO feeder cells, in DMEM, 20% foetal calf serum, 0.1 mM 2-mercaptoethanol (10). The cells were selected with G418 at 0.3 mg/ml for at least 10 days before foci were picked. In the transfection experiment to test promoter activity, E14, ES cells were used (gift from M.Hooper, Dept. Pathology). Which were grown in modified Eagles medium (Flow Labs), supplemented with 1X nonessential amino acids, 1mM sodium pyruvate, 10% foetal calf serum, 0.1 mM 2- mercaptoethanol and 10^3 units/ml LIF (ESGRO, Amstrad Corporation, Victoria, Australia). Electroporation was as described (11). G418 selection was imposed after 48 hrs, at 0.2 mg/ml, and was complete after 10 days.

* To whom correspondence should be addressed

Recombinant libraries and DNA cloning

DNA was prepared from cell lines 19/7 and 19/8 (12), and 0.05 mg of DNA was fractionated on a 1% low melting point agarose gel. DNA of the correct size fraction containing the *neo*^{R} insert was isolated using agarase digestion (13). Genomic libraries were made from this DNA in lambda L47 phage (14). The phage DNA was digested with Bam HI, ligated with the fractionated genomic DNA and packaged using Gigapack-Gold (Stratagene). Recombinant phage were selected on Mcr$^-$(P2), NM646 cells (N.Murray, Dept.Mol.Biol. Edinburgh University), and the library was amplified before replating on NM430 cells which have an *amber* mutation in the *LacZ* gene (N.Murray). Recombinant phage which contained the *SupF* gene from the NASTI construct (see Fig.1), were identified by their blue colour on X-Gal plates by complementation of the *LacZ amber* mutation. DNA was prepared from isolated recombinants (15). As the bluescribe vector is contained within NASTI, subclones containing mouse flanking DNA were quickly obtained by digesting the phage DNA with Hind III or Eco RI, self-ligating the products and using this to transform competent JM83 cells (16). Other subclones were made from DNA isolated from low melting point gels, and cloned into bluescribe vector (Stratagene).

The cDNA clone p8E2 was isolated from an 8.5 day mouse embryo cDNA library in lambda gt10 (B.Hogan, MRC NIMR Mill Hill and K.Fahrner, Biogen), using probe 8A. The cDNA was cloned into bluescribe vector, in both orientations, and subclones made using a variety of enzymes to enable the cDNA to be sequenced on both strands.

Nucleic acid hybridizations

RNA samples (0.01 mg/track), were fractionated on agarose-formaldehyde gels (15), and blotted onto Hybond-N nylon membranes (Amersham UK). DNA (0.01 mg/track), was fractionated on agarose gels and blotted (17) to Hybond-N. The nucleic acids were fixed to the nylon membrane in acordance with the manufacturers instructions and hybridized to labelled, random primed DNA isolated from agarose gels (18). Hybridizations were carried out in a buffer containing 0.5M Na Phosphate, (pH7.2); 7% SDS; 1mM EDTA at 68°C (19) and filters were given several washes of 30 mins. each in 2−0.5× SSC/0.1% SDS, at 68°C. They were then exposed to Kodak XAR-5, X-ray film, for various exposure times.

PCR analysis

DNA oligomers were made with an Applied Biosystems 381A DNA synthesizer, and were designed so that they contained a restriction enzyme site which could be used to clone the PCR product. Primer sequences from the cloned 5′ and 3′ flanking DNA were; 5′-AGAGGCATGCGGTCGTCCTCCTTC-3′ and 5′-CTTACACAGGCCTGTGGGTA-3′ from 19/7 and; 5′-GC-TGGGCAAGCTTGCTGCCC-3′ and 5′-GCTCCTGCAGAAA-CAGGAAAGG-3′ from 19/8. The amplification of the genomic DNA by PCR was as described (20). The PCR reactions were performed using 0.001 mg CCE DNA, with 50 pM of each primer, in 0.05 ml reaction mix containing 50 mM KCl/1.5 mM MgCl/10 mM Tris pH 8.25/0.2 mM dNTPs and 5 units Taq polymerase (Amersham UK). A Techne PHC-2 PCR machine was used for 30 cycles set at; 92°C for 1.5 min, 55°C for 1.5 min, 72°C for 2 min. The PCR products were then digested with Sph I/Stu I (19/7), or Hind III/Pst I (19/8), and cloned into bluescribe for sequencing.

Enzymes and sequencing

All nucleic acid restriction enzymes and modifying enzymes were from Boehringer-Mannheim Biochemicals or New England Biolabs. Sequencing was by the chain terminator method (21), using double stranded plasmids and a Sequenase kit (USB). The primers used for sequencing were the universal or M13 primers (NEB), or a synthesised DNA oligomer from the 5′ end *neo*^{R} gene.

RESULTS

ES cell transfection

ES cells were transfected by electroporation with a promoterless *neo*^{R} gene construct NASTI (Neo Activated Selection for Targeted Integration), Figure.1a.(see Materials and Methods). This construct was originally designed to mutate the mouse *Hox 2.1* gene by homologous recombination. G418-resistant foci were picked and cultured to produce nine different cell lines.

Southern blots of gel fractionated Bgl II and Bam HI restricted DNA from each cell line, probed with the *neo*^{R} gene, showed that the vector had inserted into single sites in seven of nine selected lines (data not shown). The single hybridizing bands were also of different sizes which suggests that the insertions had ocurred at different loci. These cell lines were subjected to further analysis; in these lines the *neo*^{R} gene must have trapped an active ES cell promoter.

Most insertions are in close proximity to CpG islands

Promoters for many genes lie within CpG islands. It was therefore of interest to determine whether the integration events were adjacent to such islands. CpG islands are relatively G+C rich regions, usually 1−2 kb in length, which are not deficient in the dinucleotide CpG, unlike the rest of the genome (22). CpG islands are found at the 5′ end of all known housekeeping genes, as well as some tissue specific genes, and contain transcriptional start sites (5,6). The use of rare-cutting restriction enzymes that contain CpGs in their recognition sequence is a convenient method of detecting such regions. In particular clustered sites for Not I, Sma I, Nae I, Nar I, Sac II, Bss HII, or Eag I, are diagnostic of a CpG island. Sites for these enzymes occur only rarely elsewhere in the genome and when they do are mostly blocked by methylation (23,24).

These rare cutter sites were mapped in each of the cell lines using southern blots of gel fractionated DNA, hybridized to DNA probes from within the insertion construct (Figure 1b.). In this experiment only restriction enzyme sites proximal to the insert are mapped; there may of course be multiple sites for some of these enzymes. In all cell lines except 19/10, there was a cluster of rare cutter sites flanking or adjacent to the inserted DNA. This indicates that in six of the seven lines examined the insertions have occured within or adjacent to a CpG island. In most cases the clusters of sites are within the 5′ flanking DNA or on both sides. In line 19/2 a second NASTI insert has integrated in tandem. There are island-cutter sites in the 5′ flanking DNA. The close proximity of CpG islands suggests that the promoters contained within these regions are controlling *neo*^{R} transcription. The insertion of the *neo*^{R} gene would produce a hybrid RNA from which an effective *neo*^{R} protein must be translated.

The size of the neo^R transcripts further indicate that insertions have occured close to the 5′ end of genes

A Northern blot of total RNA from each cell line, and untransfected control ES cells, was probed with the *neo*^{R} gene.

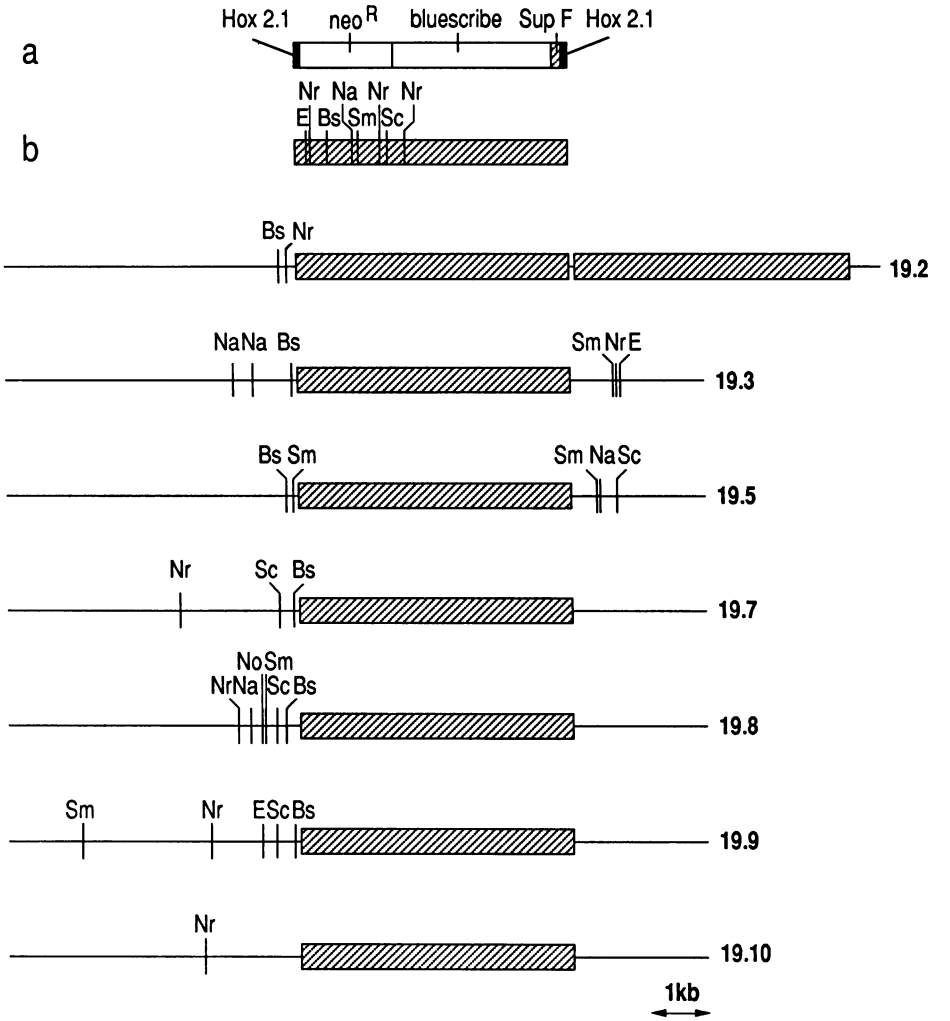

Figure 1. **(a)** The promoterless *neo*^{R} gene construct, NASTI, used in the transfection experiment. The *neo*^{R} gene has had a 5′ in-frame stop codon removed and has the HSV-Tk poly-A signal at its 3′ end (9). The construct also contains the bluescribe vector, the bacterial *SupF* gene isolated from the PiVX plasmid (50), and a 200 bp Eco RI/Bam HI fragment from the 5′ end of Hox 2.1 cDNA (51). The vector was linearized before transfection by using a unique Sac I site in the centre of the Hox 2.1 sequence, which places Hox 2.1 sequence at both ends of the construct. **(b)** A restriction map representing the inserted NASTI construct (hatched area), and flanking mouse DNA (solid line), in each of the cell lines. The vertical bars represent restriction sites for the following enzymes, Bs=Bss HII; E=Eag I; Na=Nae I; No=Not I; Nr=Nar I; Sc=Sac II; Sm=Sma I. The data was accumulated from southern blots of genomic DNA from each cell line restricted with the above enzymes, hybridised with probes from within the construct.

Each of the cell lines show hybridization to the *neo*^{R} probe (Figure.2). The intensity of the signal varies between each track. Some variation is due to unequal loading, but when compared to the actin control there is clearly a variation in activity of the endogenous promoters. The transcript sizes also vary to some extent, between approximately 1.5 and 1.8 kb. The size of the *neo*^{R} gene is 1.1 kb. The transcriptional start sites are in flanking DNA, within only a few hundred base pairs of the 5′ end of the insert. Preliminary analysis by primer extensions reveals that the transcriptional start sites in several cell lines are within 200 bp upstream of the *neo*^{R} gene and that in some cases multiple start sites lie within the *Hox 2.1* sequence of NASTI (data not shown). The close proximity of the transcriptional start sites suggests there is selection for integrations close to the 5′ end of genes possibly because longer hybrid transcripts are less likely to produce a functional *neo*^{R} protein. However, long hybrid mRNAs are known in some cases to be functional (25,26) and it may be that the chromatin at the 5′ ends of genes is more susceptible to DNA integration.

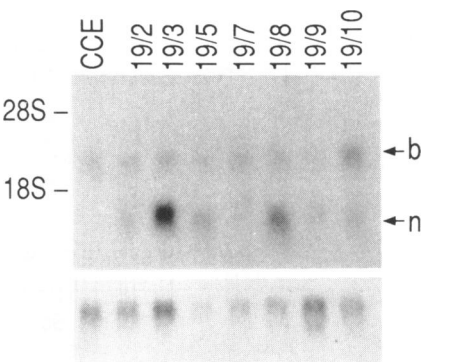

Figure 2. A Northern blot of total RNA (0.01 mg per track), from each of the transfected cell lines and of a non-transfected line, CCE and probed with the *neo*^{R} gene (upper panel), or an actin control probe (lower panel). Hybridization to the *neo*^{R} gene in each of the transfected cell lines is marked—**n**. The band marked—**b** indicates a cross-hybridizing band which is present in all of the tracks including the control non transfected line CCE.

Cloned flanking DNA from cell lines 19/7 and 19/8 confirms their CpG island nature

To enable a more detailed study to be made of the integration sites and flanking DNA, clones from cell lines 19/7 and 19/8 were selected. Both contain flanking mouse DNA 5′ and 3′ of the insert. Restriction enzyme sites within the flanking DNA were mapped. Sites for the CpG island locating enzymes confirmed the data from genomic blots and we detected additional multiple sites for some of these enzymes (Figure.3). The genomic DNA sequences at the insertion sites in 19/7 and 19/8 (see Figure.5), are over 50% G+C, with CpG roughly equal to GpC, meeting the criteria for a CpG island (22). These sequence analyses therefore confirm that DNA 5′ of the insert in both 19/7 and 19/8 is a CpG island.

Analysis of the integration points in 19/7 and 19/8 reveals short regions of homology between the genomic DNA and the construct ends

Probes from flanking 5′ DNA in 19/7 and 19/8 were used to determine if there had been any gross rearrangement of the host DNA at the integration sites. Probes 7A and 8A (see Figure.3), were used to probe southern blots of DNA from the 19/7 and 19/8 cell lines and of a non-transfected line, CCE. There is no detectable rearrangement or deletion of the genomic sequence at the integration sites. The presence of an additional Bgl II band in each of the cell line DNAs was consistent with there being a clean insertion of the construct into one chromosomal site, which increases the normal Bgl II band size by the approximate size of the integrated DNA (data not shown). To analyse the integration sites in detail, the junction sites were sequenced. These were compared to the sequences of the normal genomic DNA, determined by PCR (20), using DNA primers, 5′ and 3′ of the insertions.

Genomic sequences surrounding the insertion sites are presented in Figs.4a and 4b. A comparison between the normal genomic sequences and the DNA sequences of the integration junctions in 19/7 and 19/8, revealed no loss of DNA from the host. The points of integration are marked with an asterisk, and in both cell lines the sequence 5′-**AACTC**-3′ appears immediately 5′ of this. This sequence is complementary to 3/4 bases at the 5′ end of the insertion vector 3′-**TCGAG**-5′. There is also a short region of homology between the vector 3′ end and the genomic DNA 3′ of the integration, 4 bp in 19/7, 5′-**CAAA**-3′, and a possible single T in 19/8 (Fig.4c). It is not possible to determine which of the bases at the junction points belong to the vector and which are of host origin. This suggests that the insertion has occurred by recombination in the region of sequence similarity.

The subcloned 5′flanking DNA contains promoter elements

To determine if the DNA immediately 5′ of the *neo*^{R} gene in 19/7 and 19/8 contained promoters, the linearized subclones containing the *neo*^{R} gene, and approximately 1 kb of flanking mouse DNA, were used to transfect ES cells. If the mouse DNA contains a promoter which directs *neo*^{R} expression, the frequency of transfectants should be much higher than in a promoterless control, after G418 selection. The data (Table I), show a 100-fold higher transfection frequency for the 19/7 and 19/8 over the promoterless NASTI vector, comparable to the promoter-containing controls pSv2neo (27) and pMCneoPolyA (28). The data therefore confirm that in cell lines 19/7 and 19/8 the promoters controlling *neo*^{R} expression are contained in the genomic DNA within 1kb of the integration site.

The cloned flanking DNAs from 19/7 and 19/8 detect transcripts from endogenous genes

As NASTI has integrated downstream of a promoter, transcription of the endogenous genes may be detected using the mouse DNA flanking the insert. We therefore used DNA probes 7A and 8A (see Figure.3), to determine if endogenous gene transcription from the trapped promoters in 19/7 and 19/8 was

Table I. Results of a transfection experiment using linearized DNA vectors transfected into ES cells

DNA	Number of cells electroporated	Mean number of colonies	Frequency
NASTI	10^7	415	4.1×10^{-5}
pSV2neo	2.5×10^4	121	4.8×10^{-3}
pMCneoPolyA	2.5×10^4	168	6.7×10^{-3}
p7R1	2.5×10^4	175	7.0×10^{-3}
p8H2	2.5×10^4	148.5	5.9×10^{-3}

The p7R1 clone contains genomic DNA 5′ of the insertion in cell line 19/7, and the p8H2 clone contains genomic DNA flanking the insertion in 19/8.

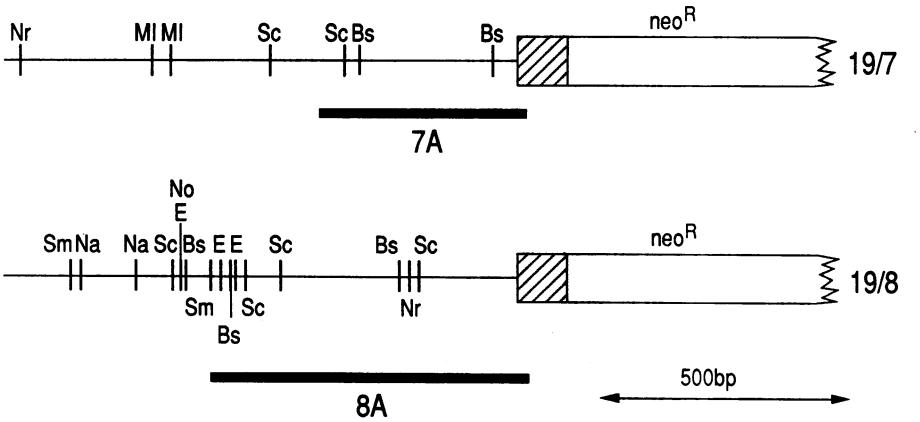

Figure 3. A map showing sites for rare-cutting restriction enzymes in the cloned mouse DNA 5′ of the insert in cell lines 19/7 and 19/8. The DNA probes 7A and 8A are indicated. These probes are from subclones p7R1 and p8H2 using Ava I digestion. They contain 25 bp of Hox 2.1 DNA from the vector end. The genomic Hox 2.1 sequence is not detected on blots due to the high stringency used. Enzymes are abrieviated as in Fig.1 and in addition, MI=Mlu I.

detectable. The probes were hybridized to Northern blots produced from total RNA from a variety of tissues.

Probe 7A, detects feint bands in most tissue types but shows strong hybridization to testis RNAs of approximately 2.0 and 3.1 kb (Figure.5a). The same bands are also detected with a 3′ probe (data not shown). Probe 8A also shows strong hybridization to testis RNA and produces strong bands of approximately 3.2 kb, 2.2 kb and a heterogeneous band of 1.2 to 1.3 kb. There is also hybridization to different RNAs in other tissues; an approximately 2.5 kb band in brain and a 2.8 kb band in muscle (Figure.5b). A cDNA clone, p8E2 isolated from an 8.5 day mouse embryo library screened with 8A, hybridizes to the small 1.2−1.3 kb transcript (Fig.5c). We suggest, therefore, that several transcripts derive from the 8A region, one of which is testis specific.

Expression has also been detected in ES cells using probe 8A on a Northern blot of polyA$^+$ RNA, and hybridizes to a band of approximately 1.8 kb (not shown). We detected very weak hybridization with 7A with poly A+ RNA from ES cells. It is

a

```
   1  TCTAGAAGAC GGCACGTGAG TGGCGAGGGG CGTGGAAAAC GCACACCTGA
  51  CGCCACTTCC TCCGCAAACC GGCAACTACT TCCCAGGCCC TCACCCTCTA
 101  CTCCGGAAGC CTTGGTTGCA GCGGAGGCTA GGGTAGCTCA CGCTCCTTTG
 151  GCATCCGAAA GGATACCAAC GGGCAGGGGC ACGACAGCCG CAGTGCCATA
 201  TTTACCGTGC CGCGGGGCCG GTGGGGAACG TCCGGGGTGC GGTTAGAAAG
 251  CCTTTCTGAA GTGTTCTCGG CTCCACGCGG TCGTCCTCCT TCAGAAACTT
 301  CAACTCCCAG AGGCAGAGGC GGTGGTTGCG CGCCCGGTTA CGGCACTTAC
                                  * → neoR
 351  GGCTTACGTC ATCGGACTGC GCTAACTCCA AACCGGAGAA AGAGGAGGAG
 401  CCTGTAGGGG GCGGGGTCTG GGCCTGGGTT TTGCGGGAAT GGTTAGTGGC
 451  TACCGCGAGG CGTTTTTCCC GGTCAAGCGT TGCATTTCTA CCCACAGGCC
 501  TGTGTAAGCT CGGGTCTCGA GGCGAAGATC GCCCACTGAA AGAAATTGAA
 551  ATGCGAAACA GACGAAATGC TCTTATTTTT TTTTT
```

b

```
   1  CCCGGGGCTG CGTGTGCAGA GCCAGCCGGC AGCCACCCGC CGCCCTCTCC
  51  GCCCGCGCTC AAGGACGTCG TGCTTTTCCT CCCAAGGACA AAGGACAATA
 101  AAGGGAGCAG GGGAGCTACT CACCGGGTCG AAGCCGGCCC CTGCGCCATC
 151  GGCTCGCTCT TCCTCGCGTG TCCAGGACGG GTCGCCCGCC GCAGCCTCCC
 201  GCCCGCGGCG CGAAGCCGCT CCCGGCCGCG CGCGACACCA GCACCGCGGC
 251  TGCCGCAGCG GAGCCACGGG CGTCGTGGGC CCGGCCTCCC CGGGCGCTCC
 301  GGCCGCGAGC TCCCACATCG CGCGCGCGGC CGCCTCGCCC CAGCCGCGGC
 351  CTCCCCGCTC TCCCAGAGCC CTCCAGGACG CCGGGGTAGC GCTCCCAGCG
 401  CAGCGGCGAG CCCACGGGCC GCGGCAGACC GCCTGACAGC TGGGCTTGCT
 451  GCTGCCTCTC CGGCCATCTT GTGCCGTGTG TGGTGTCTCT CCCGTCTCCT
 501  CGCTCGGCTT TCCTCCCTCC TCCTCCTCCC CCGGCCCACA CACCACACAA
 551  CATGGCCTCT TCGCTGCGGC GGGAGCTGCG GCGGCGGCTA GAGCTCGCGG
 601  CGCTCCAGCT CGCTCCCCTC CCCCAACCTC CCGCCTGCAG CTCCGGGACC
 651  CCCAGTGCGC ACGCGCGCCG ACACCCGCCC GCGGGCGCCA TTGGCTGGGC
 701  TAGCTTCCTG CCCTCGCTTC CCGGAGCGGC ATCACGCACT TCGTTCGTAC
 751  CTTCCTCCCG CCGCAACTGC ATACTATTGG TCAGTTCTGA ACCCACCGCT
 801  TCCATGTTCT CAGCACCAAT AGGGCTTATT TAATGCCAAT CATAGCCCGT
                                  *→neoR
 851  GGGTTTCAGG AGAGAAGTCC TGACCTGCGA TTGGCAGATA GTGAAACTCT
 901  GGGTTTGCGG GGAGAGACGT AACTTTCTGC TGCAGTGCGG ACCCTCGAAG
 951  GTAGTTTGTG TACGGACGAG CCATTAAAAG CTAACCTTTC CTGTTTCTGA
1001  CTTGAGATTT TTTTCGTCCA CAAAGGGAAG AGGAAGTGAT AGAGCG
```

c

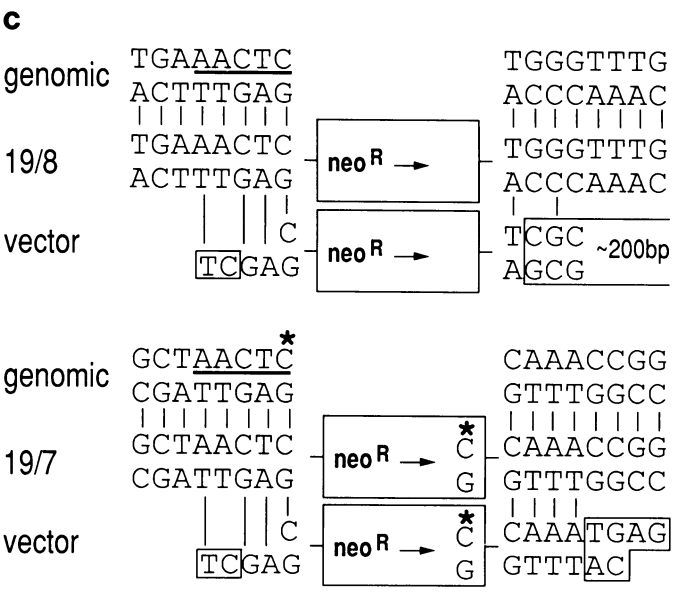

Figure 4. **(a)** The genomic DNA sequence flanking the insertion site in 19/7. The asterisk represents the point of insertion and the direction of *neo*^{R} transcription is indicated. **(b)** The genomic DNA sequence flanking the insertion site in 19/8. The asterisk represents the point of insertion and the direction of *neo*^{R} transcription is indicated. The underlined sequence is the same as the 5′ end of the p8E2 cDNA clone and continues upto a consensus splice site (29), on the opposite strand. A TATA-like box is double underlined, as is the consensus CAAT box (52). GC boxes are enclosed in brackets. **(c)** Sequence homologies present at the integration sites in 19/7 and 19/8. Homologous bases between the normal genomic DNA, the transfected cell line DNA and the insertion construct are shown as vertical lines. The asterisk in 19/8 denotes a base which could be on either side of the genomic breakpoint. The 5 bp sequence, AACTC, which is present at both integration points, is underlined. The boxed area represents the NASTI construct and the direction of *neo*^{R} transcription is indicated. The vector DNA in both cases remains intact except for the boxed sequence at the construct ends which have been deleted on insertion.

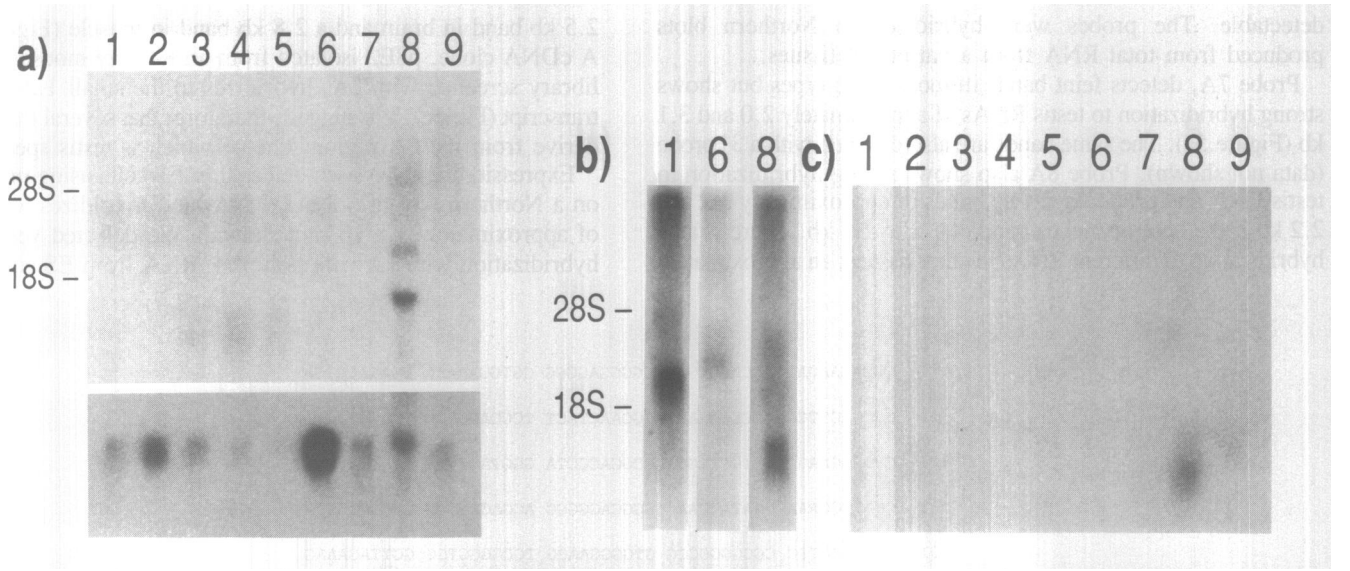

Figure 5. (a) A Northern blot of total RNA from various tissues and hybridized to probe 7A (see Fig.2). (1=Brain; 2=Heart; 3=Kidney; 4=Liver; 5=lung; 6=muscle; 7=ovary; 8=testis; 9=thymus). The size of the 18S and 28S rRNAs are indicated. The lower panel shows the same blot hybridized to a GAPDH probe (53). Note this probe detects high expression in muscle. (b) A Northern blot showing tissues which hybridize to probe 8A (see Fig.2). Tissue types are numered as above. (c)The same Northern blot as in Fig.5a hybridized to the cDNA clone p8E2. The size of the 18S and 28S rRNA bands are indicated. Tissue types are numbered as above.

possible that the insertion in 19/7 has trapped a promoter which does not give rise to a stable transcript in ES cells or has activated a promoter normally inactive.

We therefore detect transcription of DNA which flanks the insert in both cell lines from what appear to be non-housekeeping genes with high expression in adult testis.

The promoter within the 5′ flanking DNA of cell line 19/8 has a bi-directional activity

The sequence of the cDNA clone, p8E2, which is contained in 19/8 is shown by underlining in Figure.4b. The end of the sequence match to 19/8 corresponds with a consensus 5′splice site (29), and the cDNA sequence continues for another 558 bp beyond this. The cDNA also has a polyadenylation signal AATAAAA and a poly-A tail 13 bp downstream of this. The splice site and polyA tail indicate that the mRNA from which the cDNA derives is transcribed in the opposite direction to the *neo*^{R} gene. All three reading frames are open in the cDNA, for at least 647 bp. We do not know which is translated but none show significant amino acid similarity to any sequence in the databases. Between the cDNA and the NASTI insertion in 19/8 there are a series of potential transcription factor binding sites. These include a TATA-like box in both orientations, and a CCAAT box appropriately distanced from the TATA box to regulate the p8E2 mRNA. There are five GC boxes, GGGCGG, which are potential binding sites for the transcription factor Sp1 (30), and could control expression in either or both directions. DNA binding sites for Sp1 are known to be important in the regulation of transcription in many genes, and are commonly found within CpG islands (6). The binding of Sp1 may regulate transcription in the presence or absence of either a TATA or CAAT box and may direct transcription of the *neo*^{R} gene in 19/8. Other bi-directional promoters have been described (31,32,33,34), and may frequently occur in CpG islands (33). We are at present analysing the bi-directional promoter from 19/8 in more detail.

DISCUSSION

We present an analysis of several G418-resistant cell lines produced from a single transfection experiment using a promoterless *neo*^{R} gene construct, NASTI. We have used the method as a promoter trap to identify and mutate members of a particular class of genes, those which are active in ES cells, and to analyse the integration sites to determine the mechanisms of non-homologous recombination.

The use of promoter and enhancer traps as a means of cloning genes and regulatory regions has been used in rat cells (35), and by microinjection of transgenes into mouse embryos (36,37). DNA transfection and retroviral infection of EC and ES cells has also been used for this purpose (38,39,40). Here we have shown that in seven cell lines with single insertion sites, six have occurred close to or within a CpG island. As our experiment was designed to select for active promoters, it was not unexpected that such regions would be targetted, as they contain promoters for most genes (5,6). However the high frequency of targetting genes with CpG islands may indicate that there are a larger number of such genes active in ES cells. Alternatively the chromatin of CpG islands may be more susceptible to DNA integration than non-island chromatin. We are currently examining this latter hypothesis in more detail.

It has been previously suggested that expressed genes with an opened chromatin conformation are likely targets for integration (41), and it has been shown that CpG islands have such an altered structure (7). DNA in these open regions may be transiently exposed during normal cellular processes such as transcription and replication and allow a high level of recombination with exogenous DNA. Our data are consistent with this hypothesis and could also account for the high frequency of homologous recombination seen with some targetted genes such as *HPRT* (42), and *N-myc* (25).

Two integration sites analysed have the same sequence, 5′-**AACTC**-3′, in the genomic DNA. Interestingly this sequence encompasses the trinucleotides CTC on one strand and GTT on

the other. Both of these have been found to be significantly overrepresented in a compilation of non-homologous integration sites of transfected linear DNAs and may represent sites for DNA topoisomerase I (43). The compilation also shows a significant occurrence of CTC at position −1 and GTT at position +5 of the breakpoints, which is where they occur in cell lines 19/7 and 19/8.

There has been speculation that many non-homologous recombinations take place at Topo I sites (see review 44). Furthermore, the integration sites in 19/7 and 19/8 also show a 4 bp match with the 5′ end of the transfected DNA construct and a 4 bp and 1 bp match respectively, with the 3′ end. Such short regions of identity, between 1 and 5 bp, have been found at a variety of 'illegitimate' recombination junctions in mammalian cells (45,46,47). We propose therefore that Topo I may initiate recombination by causing a nick in one, or both strands, and that short regions of homology near the breakpoint are exposed to recombination with the transfected DNA. This would account for the integration of the vector into the mouse genome in 19/7 and 19/8, without the addition of extra nucleotides at the junction, or rearrangements commonly found at other integration sites (48).

The results also show that the DNA within 1 kb of the *neo*^{R} gene in 19/7 and 19/8 contains promoter elements which direct *neo*^{R} transcription, and these promoter elements are also responsible for the transcription of endogenous genes, which have been detected by Northern blot analysis. The promoter region in 19/8 has been shown to have bi-directional activity because it is able to transcribe *neo*^{R} in one direction and the p8E2 transcript in the other. Several CpG islands show bidirectional promoter activtiy (30,31,32,33). Recently Johnson and Friedman (49), tested two specific CpG island promoters, human *HPRT* and *PGK*, and found a limited bidirectional activity. We are currently screening cDNA libraries to characterise the genes transcribed at the insertion site of 19/7 and 19/8.

In summary, we have shown that electroporation allows integration of DNA in a manner which minimizes rearrangement of both vector and genomic DNA. By using a promoterless construct we select for integration close to active promoters. The frequency of insertions close to CpG islands suggests that these are not only indicative of the 5′ end of genes, but also that these regions may be more susceptible to DNA integration. We propose that non-homologous integration of DNA occurs by preferential Topo I cleavage at open chromatin, such as at CpG islands, followed by recombination through short regions of sequence similarity.

This method produces cell lines which can be used to generate chimaeras with mutated genomes, marked by the insert at CpG islands. These may provide a link between a physical map of the genome, and a mutational map.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council and the Lister Institute of Preventative Medicine. We would like to thank Adrian Bird, Nick Hastie and Helen Sutherland for their comments on the manuscript, Doreen Chambers for making the DNA oligomers, and our excellent photographic dept. for producing the figures. IJJ is a Lister fellow.

REFERENCES

1. Evans,M.J., Kaufman,M.H. (1981) *Nature* 292;154−156.
2. Martin,G.R. (1981) *Proc.Nat.Acad.Sci.* **78**;7634−7636.
3. Robertson,E.J. (1986a) *Trends Gen.* 2;9−13.
4. Capecchi,M.R. (1989) *Science* 224;1288−1292.
5. Bird,A.P. (1987) *Trends Genet.* **3**;342−347.
6. Gardiner-Garten,M., Frommer,M. (1987) *J.Mol.Biol.* **196**;261−282.
7. Tazi,J., Bird,A. (1990) *Cell* 60;909−920.
8. Robertson, E.J., Bradley,A., Kuehn, M., Evans,M. (1986b) *Nature* **323**;445−448
9. Dorin,J.R., Inglis,J.D., Porteous,D.J. (1989) *Science* 243;1357−1360.
10. Robertson,E.J.(1987) In*Teratocarcinomas and embryonic stem cells; A practical approach.* (ed. E.J.Robertson) IRL Press, Oxford and Washington DC. pp**71−112**
11. Thompson,S.,Clarke,A.R.,Pow,A.M.,Hooper,M.L.,Melton,D.W.(1989) *Cell* 56;313−321.
12. Lovell-Badge,R.H. (1987) In *Teratocarcinomas and embryonic stem cells: a practical approach,* (ed.E.J.Robertson),pp. 153−182. IRL Press.
13. Burmeister,M., Lehrach,H. (1989) *Trends Genet.* **5**;41.
14. Loenen,W.A.M., Brammar,W.J. (1980) *Gene* 10;249−259.
15. Maniatis,T., Fritsch,E., Sambrook,J. (1982) *Molecular Cloning: A laboratory manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
16. Messing,J, (1979) *Recomb. DNA Tech.Bull.* **2(2)**;43−48.
17. Southern,E. (1975) *J.Mol.Biol.* **98**;503−517.
18. Feinberg,A.P., Vogelstein,B. (1984) *Anal.Biochem.* **137**;266−267
19. Church,G., Gilbert,W. (1984) *Proc.Nat.Acad.Sci.*81;1991−1995.
20. Saiki,R., Gelfano,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. Erlich,A.H. (1988) *Science* **239**;487−491.
21. Sanger,F., Nichen,S., Colsen,A. (1977) *Proc.Nat.Acad.Sci.* **74**;5463−5467.
22. Bird, A.P. (1986) *Nature* **321**;209−213.
23. Lindsay,S., Bird,A.P. (1987) *Nature* **327**;336−338.
24. Bird,A.P. (1989) *Nucleic Acid Res.*17;9485.
25. Charron,J., Malynn,B.A., Robertson,E.J., Goff,S.P., Alt,F.W. (1990) *Mol.Cell.Biol.* **10**;1799−1804.
26. Vaeck,M., Reynaerts,A., Hofte,H., Jansens,S., De Beuckeleer,M., Dean,C., Zabeau,M., Van Montagu,M.,Leemans,J. (1987) *Nature* **328**;33−37
27. Southern,P.J.,Berg,P. (1982) *J.Mol.Appl.Genet.* **1**;327−341.
28. Thomas,K.R., Capecchi,M.R. (1987) *Cell* **51**;503−512.
29. Shapiro,M.B., Senapathy,P.(1987) *Nucleic Acids Res.*15;7155−7174.
30. Kadonaga,J.T., Jones,K.A., Tjian,R. *Trends Biol.Sci.* **11**;20−23
31. Farnham,P.J., Abrahms,J.H.,Shimke,R.T. (1985) *Proc.Nat.Acad.Sci.* **82**;3978−3982.
32. Mitchell,P.J., Carothers, A.M., Han, J.H., Harding, J.D., Has, E., Veniola, L., Chasin,L.A. (1986) *Mol.Cell.Biol.* **6**;425−440.
33. Lavia.P., Macleod,D.T., Bird,A. (1987) *EMBO J.* **6**;2773−2779.
34. Poschl,E., Pollner,R., Kuhn,K. (1988) *EMBO J.* **7**;2687−2695.
35. Fried,M., Griffiths,M., Davies,B., Bjursell,G., La Matia,G., Lania,L. (1983).*Proc.Nat.Acad.Sci.* **80**;2117−2121.
36. Allen,N.D., Cran,D.G., Barton,S.C., Hettle,S., Reik,W., Surani,M.A. (1988) *Nature.* **333**;852−855.
37. Kothary,R., Clapoff,S., Brown,A., Campbell,R., Peterson,A., Rossant,J. (1988) *Nature* **335**;435−437.
38. Bhat,K., Burnley,M.W., Hamada,H. (1988) *Mol.Cell.Biol.* **8**;3251−3259.
39. Gossler,A., Joyner,A.L., Rossant,J., Scarnes,W.C. (1989) *Science* **244**;463−465.
40. Peckham.I., Sobel,S., Comer,J., Jaenisch,R., Barklis,E. (1989) *Genes Dev.* **3**;2062−2071.
41. Jaenisch,R. (1988) *Science* **240**;1468−1474.
42. Mansour,S.L., Thomas,K.R., Capecci,M.R. (1988) *Nature* 336;348−352.
43. Konopka,A.K. (1988) *Nucleic Acids Res.* **16**;1739−1758.
44. Champoux,J.J., Bullock,P.A. (1988) In *Genetic Recombination*,(eds. Kucherlapati,R., Smith,G.R.), pp. 655−666.Am.Soc.Microbiol.
45. Roth,D.B., Wilson,J.H. (1986) *Mol.Cell.Biol.* **6**;4295−4304.
46. Bullock,P. (1985) *Science* 230;954−958.
47. Murname,J.P., Young,B.R. (1989) Gene 84;201−205.
48. Roth,D., Wilson,J. (1988) In *Genetic Recombination* (eds. R.Kucherlapati and G.R.Smith), pp. 621−653. Am.Soc.for Microbiol.
49. Johnson,P., Friedman,T. (1990) *Gene* **88**;207−213.
50. Seed, B. (1983) *Nucleic Acid Res.***11**;2427−2445.
51. Krumlauf,R., Holland,P.W.H., McVey,J.H., Hogan,B.L.M. (1987) *Development.***99**;603−617.
52. Bucher,P. (1990) *J.Mol.Biol.* **212**;563−578.
53. Edwards,Y.H., Lloyd,J.C., McMillan,S.L., Benham,F.J. (1985) *Mol.Cell.Biol.* **5**;2147−2149.